

# Scaling and Controlling Server Energy and Performance

Erez Zadok++

ezk@cs.stonybrook.edu

Filesystems and Storage Lab  
Stony Brook University  
Computer Science Department



8/8/2011

HEC-FSIO 2011

## Claim #1

- Software is wasteful
  - ◆ ...and dumb

[ACM SYSTOR '09]



8/8/2011

HEC-FSIO 2011

2

## Compression Study

- Three dimensions of the evaluation:
  - ◆ Hardware type (testbed)
    - Server, desktop, and laptop machines
  - ◆ Compression algorithm
    - gzip, lzop, bzip, compress, and ppmd
  - ◆ Input data: Text, binary, random, zeros
- Q: is it worth compressing?
- Results: It depends
  - ◆ Best: save 10x energy+performance
  - ◆ Worst: waste 200x

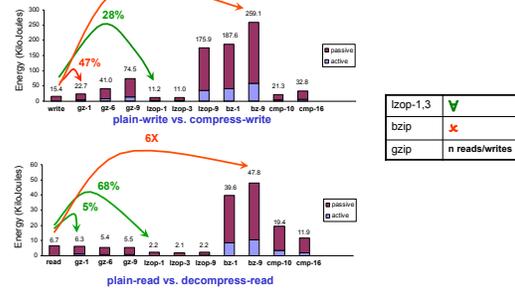


8/8/2011

HEC-FSIO 2011

3

## Server Energy: Text File



8/8/2011

HEC-FSIO 2011

4

## Server Break-Even Values (E, T)

Tool	Text	Binary	Zero	Random
gz-1	20.2 , 2.8	✗	✓	✗
gz-6	19.7 , 10.9	✗	✓	✗
gz-9	49.0 , 24.9	✗	✓	✗
lzop-1	✓	2.1 , 0.8	✓	✗
lzop-3	✓	2.1 , 1.0	✓	✗
lzop-9	35.4 , 26.8	172 , 130	5.4 , 3.6	✗
bzip-1	✗	✗	0.28 , ✓	✗
bzip-9	✗	✗	1.3 , 0.2	✗
c-10	✗	✗	✓	✗
c-16	✗	✗	✓	✗



8/8/2011

HEC-FSIO 2011

5

## Claim #2

- Everything affects performance and energy
  - ◆ ...much more than you thought.

[FAST'10, TOS'10]



8/8/2011

HEC-FSIO 2011

6

## Workload Study

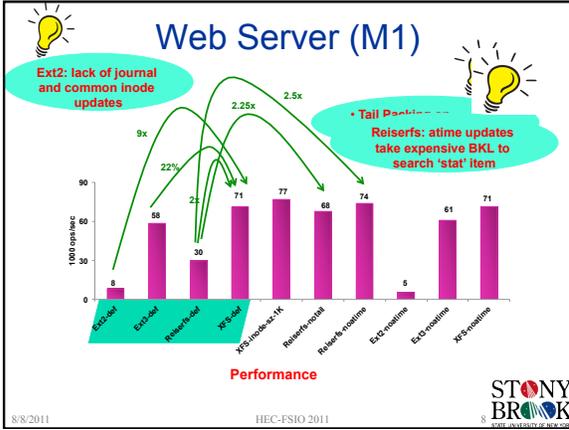
- **Server Workloads (4)**
  - ◆ Web, Database, File server, Mail
  - ◆ FileBench emulated workloads
- **File Systems (4)**
  - ◆ **Type:** Ext2, Ext3, ReiserFS, XFS
  - ◆ **Mount Options:** noatime, notail, journal=<modes>
  - ◆ **Format Options:** inode size, blocksize, allocation/block group count.
- **Hardware (2)**

8/8/2011

HEC-FSIO 2011



## Web Server (M1)



8/8/2011

HEC-FSIO 2011



## File System Selection Matrix (M1)

Workload	Best File System (Combination)	Improvement Range (compared to all default FS)	
		Ops/sec	Ops/joule
Web Server	XFS (inode-size-1K)	8% - 9.4x	6% - 7.5x
File Server	ReiserFS (default)	0% - 1.9x	0% - 2.0x
Mail Server	ReiserFS (notail)	29% - 5.8x	28% - 5.7x
Database Server	XFS/Ext3 (BLK-2K)	2.0 - 2.4x	2.0 - 2.4x

8/8/2011

HEC-FSIO 2011



## Claim #3

- We often don't know what to benchmark in the storage stack
  - ◆ ...or we do a poor job at it

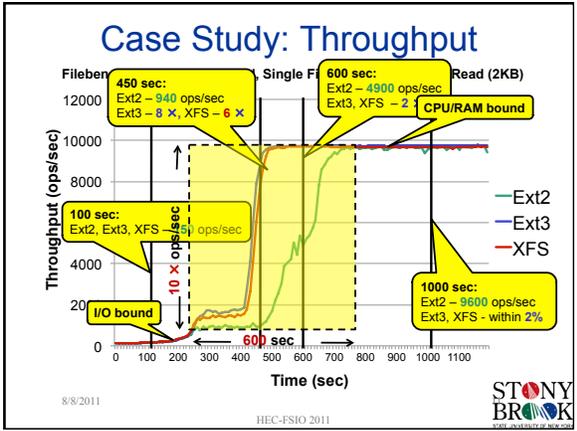
[ACM TOS '08, HotOS '11]

8/8/2011

HEC-FSIO 2011



## Case Study: Throughput

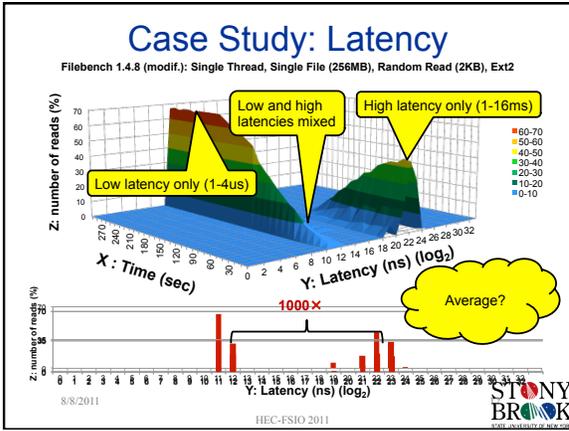


8/8/2011

HEC-FSIO 2011



## Case Study: Latency



8/8/2011

HEC-FSIO 2011



## FileBench

- Official maintenance as of May'11
- Autotools, ports, bugfixes
- OSprof (2D/3D) histograms
- Energy measurements
- More random distributions, extensible
- Ongoing:
  - ◆ Generate multi-modal distributions
  - ◆ Convert traces into workload models
  - ◆ Generate compressible/dedupable patterns

8/8/2011

HEC-FSIO 2011



## Claim #4

- You can scale by orders of magnitude
  - ◆ ...but you have to “think” differently

[SOCC '11]

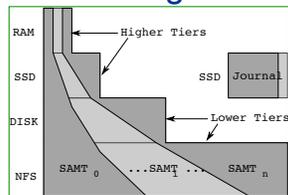
8/8/2011

HEC-FSIO 2011



## Tablet Server Storage

- Future: heavy indexing workloads
- SAMT: Sorted Array Merge Trees
  - ◆ Can merge efficiently
- Transactional KV store
- Hot items percolate up
  - ◆ Colder ones down
- Scales better than B-trees
- 10–1000x better than BDB, MySQL, XFS/Ext3, best log-structured index

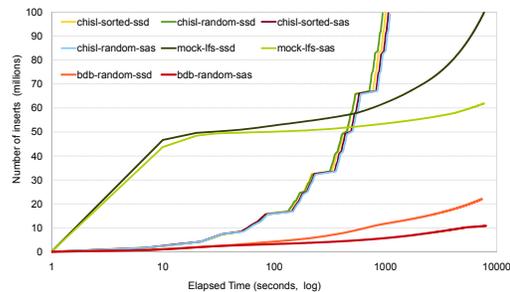


8/8/2011

HEC-FSIO 2011



## Insertion Performance



8/8/2011

HEC-FSIO 2011



## Claim #5

- We don't use good algorithms
  - ◆ ...or data structures
- “systems” vs. “theory”?

[HotStorage '11]

8/8/2011

HEC-FSIO 2011



## New Algs & Data Structs

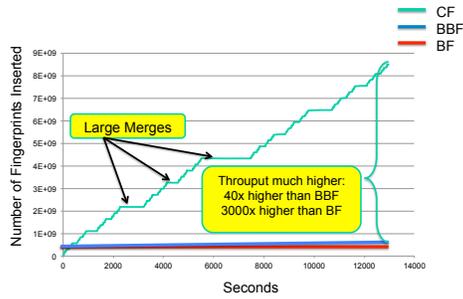
- **Cascade Filter (CF)**, a BF replacement optimized for fast inserts on Flash
- Our performance
  - ◆ We do 670,000 inserts/sec (40x of other variants)
  - ◆ We do 530 lookups/sec (1/3x of other variants)
- We use **Quotient Filters (QF)** instead of Bloom Filters
  - ◆ They have better access locality
  - ◆ You can efficiently merge two QFs into a larger QF (w/ same FP rate)
- We use **merging techniques** to compose multiple QFs into a CF

8/8/2011

HEC-FSIO 2011



## Insertion Throughput



8/8/2011

HEC-FSIO 2011



## Claim #6

- Optimizing and controlling multiple-dimensions is hard
  - ◆ ...revisit control theory?

[ACM SYSTOR '11, ERSS '11]

8/8/2011

HEC-FSIO 2011



## Intelligent Compression?

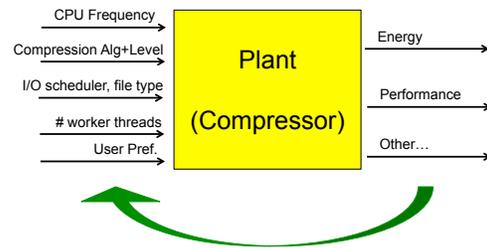
- Balance Energy, Performance, Reliability, \$\$\$, etc.
  - ◆ User tradeoff inputs
- Use Control Theory
  - ◆ Self-adaptation
- Problems with traditional controllers
  - ◆ Assumes linear, stable behavior, low st.dev
  - ◆ Assumes single inputs/outputs
  - ◆ Assume numeric, meaningful inputs

8/8/2011

HEC-FSIO 2011



## Desired Plant Model



8/8/2011

HEC-FSIO 2011



## System Identification Problems

- Nonlinearity
- Instability
- Multi-dimensionality
  - ◆ CPU Frequency
  - ◆ I/O Schedulers
  - ◆ File Types
  - ◆ Disk Types
  - ◆ Compression Algorithm + Level
- Non-numeric labels

8/8/2011

HEC-FSIO 2011



## Techniques Being Investigated

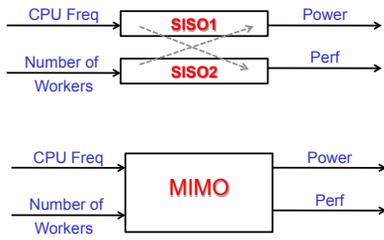
- Data Mining, HMMs
- Visual Analytics
- Multi-Dimensional Scaling
- Hierarchical controllers
- Segmentation
- Multiple-Input Multiple-Output models
- ...

8/8/2011

HEC-FSIO 2011



## Models



MIMO model and two SISO models

8/8/2011

HEC-FSIO 2011



# Q&A

**Erez Zadok++**

[ezk@cs.stonybrook.edu](mailto:ezk@cs.stonybrook.edu)

*Filesystems and Storage Lab*  
Stony Brook University  
Computer Science Department



8/8/2011



HEC-FSIO 2011

